A Discourse Marker Tagger for Spanish using Transformers

Etiquetador automático de Marcadores Discursivos mediante Transformers

Ana García Toro, Jordi Porta Zamorano, Antonio Moreno-Sandoval
Universidad Autónoma de Madrid
{ana.garciatoro, jordi.porta, antonio.msandoval}@uam.es

Abstract: We present an automatic discourse particle (DM) tagger developed using manual annotation and machine learning. The tagger has been developed on a dataset of financial letters, where human annotators have reached an 0.897 agreement rate (IAA) on the indications of a specific annotation guide. With the annotated dataset, a prototype has been developed using the pre-trained Transformers, adapting it to the task (fine-tunning), reaching an F1-score of 0.933. An evaluation of the results obtained by the tagger is included.
Keywords: Discourse Markers, Spanish, fine-tuning Transformers.

Resumen: Presentamos un etiquetador automático de partículas discursivas (DM) desarrollado mediante etiquetado manual y aprendizaje automático. El etiquetador se ha desarrollado en un *dataset* de cartas financieras. Las anotadoras humanas han alcanzado un 0,897 de tasa de acuerdo (IAA) sobre las indicaciones de una guía de anotación específica. Con el *dataset* anotado se ha desarrollado un prototipo usando modelos de Transformers pre-entrenados adaptándolos a la tarea (*fine-tuning*) con un F1 de 0,933. Al final se da una evaluación de los resultados obtenidos por el *tagger*.
Palabras clave: Discourse Markers, Spanish, fine-tuning Transformers.

1 Why a Discourse Marker Tagger?

1.1 What is a DM?

Discourse Markers (DMs) are a large and heterogeneous group of invariable linguistic units that constitute intra- and supra-speech links for textual cohesion and coherence. Their primary function is to mark and define the relationship between the parts of the speech and to guide the inferences of the discourse from a procedural approach (Zorraquino and Portolés, 1999; Pons, 2000; Montolío, 2001; Briz et al., 2008; Fuentes 2009; Landone, 2012). Among the inferences that DMs guide are structuring information (1), counter-arguing opposite ideas (2), adding information (3), focusing on relevant nuances (4), introducing new arguments or statements (5), and reinforcing elements of the speech. In any case, they cohere and structure the discourse to be satisfactorily understood.

Here are some examples from our FinT-esp financial corpus (Moreno-Sandoval et al., 2020):

(1) *En 2015 nuestros dos objetivos fundamentales son:* ***por un lado****, seguir mejorando la franquicia comercial para estar en disposición de ganar cuota en una economía en crecimiento (...)*
(2) ***Por el contrario****, los resultados por operaciones financieras caen un 16% afectados por la volatilidad del mercado*
(3) ***También*** *es destacable el aumento del crédito, por primera vez desde 2008, por el impulso de empresas y pymes,* ***así como*** *el fuerte crecimiento en la producción de nuevas hipotecas*
(4) *Se trata de un resultado impulsado* ***principalmente*** *por el impacto de los 605 millones de euros de plusvalía obtenidos con la venta del 34% de Cellnex Telecom*
(5) ***Con respecto a*** *la tecnología, estamos invirtiendo fuertemente para ser más eficientes y abaratar los procesos*

Following previous definitions of DMs for written texts, we tried to find the best description

for our financial corpus. In financial texts, the writer's aim of using them is that the reader arrives at a particular interpretation of the utterances through certain inferences (persuasion). DMs are, therefore, essential keys to financial discourse.

It is important to consider that most of the definitions given by the literature have been applied to oral and written texts in general. For this reason, they are broader and less accurate definitions to include all DMs, despite their very different characteristics. They all have a semantic feature in common (with a few exceptions): DMs are characterised by a lack of referential or propositional content. Some authors (Llamas et al., 2010) have focused their taxonomy on one discourse's type or genre. In the case of Llamas et al. (2010), they have classified DMs in academic texts, while others have done it for oral discourse (Briz et al., 2008). In any case, their definition and classification have been a controversial field for scholars (Loureda and Acín, 2010) because each author considers different elements, concepts, and properties to categorise the DMs. Following previous definitions of DMs for written texts, we tried to find the best definition according to our financial, written, and formal discourse. Discourse Markers are a large and heterogeneous group of linguistic units that constitute intra- and supra-speech links for textual cohesion and coherence. The aim of using these elements by the sender of the text is that the interlocutor arrives at a particular interpretation of the utterances through certain inferences. Given the lack of studies on discourse markers in financial narratives, we provide this definition.

It should be noted that DMs are not grammatical elements, nor do they have a conceptual meaning. In other words, they do not have a defined place in the syntax; they act at various levels of discourse (depending on their function) and do not provide lexical information. DMs give information on how ideas in discourse are related.

Because of the diversity of the original categories (adverb, preposition, conjunction) and their behaviour in discourse, it isn't easy to establish a boundary between what is and what is not a DM. This makes the functional category of DMs a semi-open category. Let us say that not everything can function as a marker, but that, as the name suggests, they are words that mark discourse, and guide certain inferences. Even so, the list of DMs is neither closed nor defined according to certain established features; in fact, in this work we have been able to verify that the influence between languages can generate new incorporations of these elements into a language, as has happened in Spanish with *adicionalmente*, which we think comes from the English *additionally*. So, which words can mark the discourse and perform the functions of a DM? The complexity of this DM-tagger is, precisely, that we are dealing with a task that is difficult for linguists to define.

1.2 Why a DM-Tagger?

To understand the role of these particles, which may or may not be integrated into the sentence, we must know they form an essential part of it. The function of a DM-tagger goes beyond distinguishing them in context. It allows the reader to understand how discourse structures work, their distribution, and their purpose. Specifically, regarding the issue at hand, they can also give clues about the company's financial results. DMs seem to show imperceptible inferences in the text, guiding our thoughts and beliefs about the company. In short, DMs guide discourse and are tools of persuasion and manipulation, which are of great interest to speakers in business discourse.

A DM-tagger and its automatic annotation may involve the introduction of an objective measuring instrument that can resolve theoretical discussions. This tagger aims to reduce the inherent subjectivity that results from studies carried out with introspective methods and with which they end up doing manual and controversial classifications.

This tagger, applied to financial discourse, can be a tool applicable to any other type of discourse for identifying DMs. Knowing its distribution, functions, or behaviour are the first steps to a better understanding of the structure and construction of the discourse and, above all, to see how the discourse can be transmitted, manipulated, or lied through. DMs help study all discourse structures since they are part of them, whether in the sentence or outside it.

Of course, as we said, they are essential for discourse comprehension, but not indispensable; that is to say, they are a great help to the reader, reducing errors of interpretation and textual ambiguity.

Besides theoretical consequences, this work has several practical applications such as

discourse segmentation, information extraction, automatic summaries, or machine translation.

1.3 Previous Work on DMs Tagging

The main problem for the classification of DMs, and even more so for automatic classification, is the lack of consensus among scholars as to what is a DM and in which contexts an item is considered a DM, and in which are not. In our case, we also must consider the annotator bias in shaping the annotation guide. Proposals have been made since the 1990s to detect and systematise DMs. But neither the first investigations nor those carried out at the beginning of the century achieved results with a high percentage of precision and accuracy, due, once again, to the lack of consensus that exists when it comes to defining the units that do or do not fall into this group. Alonso et al. (2002), and, subsequently, Muller et al. (2016) undertook the construction of a computational lexicon of previously hand-coded DM, using the clustering technique to group markers (mostly connectors) that shared syntactic contexts or, in other words, that had similar behaviour. In both cases, the categories collected were hardly verifiable in the corpus because they had not been compiled based on an actual text but on a predefined lexicon.

Hernán et al. (2017) present a proposal for automatic induction of classifications of DMs that behave parenthetically (at the margin of the sentence separated by punctuation marks). They used a parallel corpus to automatically induce DMs categories according to the similarity between Spanish and English elements, without any prior annotation, which has not been done to date. In their update, Hernán and Nazar (2018) achieved high DM/non-DM decision accuracy.

Lastly, Rogelio Nazar, following his previous work in DMs identification and classification in Hernán et al. (2017) and Hernán and Nazar (2018), presents in Nazar (2021) a methodological proposal for the automatic induction of a multilingual taxonomy of DMs through parallel corpora. Using statistical calculations, he separates DMs from the rest of the units because, due to their low amount of referential information, they act "randomly" when grouped with other units in the text (as opposed to lexical units, which syntactically behave in a regular way). Then, once the DM candidates are selected automatically, they are aligned in pairs with their equivalents in other languages without any human intervention. In this way, the DMs in one language and similar DMs in others belong to the same category because they behave similarly. At first, these categories into which the DMs are grouped are not labelled with any name, but once they already contain a considerable number of DMs, the terminology followed is the one contributed a couple of decades ago by Zorraquino and Portolés (1999). Finally, this clustering technique is used to obtain and classify new units from these categories.

They collected 2636 items divided into 70 categories, which human annotators then reviewed. The review revealed that the model had 95% accuracy in the languages chosen for the experiment: English, Spanish, Catalan, French and German, except the latter, with 84%, probably due to the morphological characteristics of this language.

One of the disadvantages of this type of methodology is the 100% automatic selection and classification of DMs, in which the context is not considered. The information around a functional element such as these is important because they function as DMs in some contexts, while in others do not. Discerning contexts in which a DM functions seems to be a task that requires previous human annotation.

As for us, we provide an approach to the study of DMs in financial narrative from an actual perspective of their behaviour and distribution. Our analysis is based on a corpus annotated and contrasted by two annotators. However, it is still subject to certain underlying theoretical conjectures of linguistic introspection and the foundations laid by experts.

2 Dataset

The documents used in this research belong to the financial domain, characterised by a specialised language and a particular communicative exchange. The interlocutors are usually specialists in the field of finance and business. We will focus on letters written by managers to their investors (see 2.1).

The financial narrative in Spanish, in contrast to English, presents an excessively technical discourse with a significant contribution of English terminology (Mateo 2007, Vargas and Carbajo 2021). Mateo (2007) goes so far as to state that financial texts in Spanish are obscure and complex and that the reading of the financial

press is so dense that its content is not within reach of the non-specialist reader.

On the other hand, the exhortative communicative function predominates in the particular documents considered here (see 2.1) The sender intends to convince the receiver of their company's benefits so that he/she invests in it.

For this reason, we have found it to be a good testing ground for the use of DMs in argumentation.

2.1 Letters to Shareholders Corpus

Letters to Shareholders (LTS) is a sub-type of the financial narrative genre. They are the summaries that appear in companies' annual reports. It has recently attracted some interest in the NLP field: El-Haj et al. (2019), Moreno et al. (2019), and Bel et al. (2021).

Gisbert (2021) describes the two argumentative strategies used by managers:

a. Emphasising the company's good results, thanks to good management.

b. Hiding negative information that affects the expectations and reputation of the company and its managers.

For the work presented in this paper, we have chosen a subset of 397 letters in Spanish, with a length of 462,189 words and 16,800 sentences.

2.2 Annotated Dataset

Linguists have manually annotated the LTS corpus with DM tags in different stages, explained in section 3.2. In the complete annotation process (see section 3.2), 3170 DMs have been annotated, which appear in a total of 6432 sentences, containing a total of 154219 tokens. The distribution in each phase is shown in Table 2 (see section 4).

3 Annotation Process

3.1 Guidelines

The task of our annotation guide was to collect only the Discourse Marker (DM) category; this is to say, annotating only terms that were discourse markers.

Multi-label annotation to account for sub-categories has not been handled in this phase and is left for further work since we only wanted to approach DMs annotation. Revising previous work on DMs classification, we noticed the classification criteria' issues. Defining a Discourse Marker and its limits as a functional element is complex enough, considering that the classification used for a type of text is useless for others (notice the differences between all the DMs used in oral discourse that are never used in written texts). In addition, researchers disagree with groups of DMs and their sub-groups. These were the reasons why, for this study, as a first step, it was decided to annotate the binary task of DM/non-DM.

We train the model with a more significant number of DMs, although some of them may be under-represented. We prioritised coverage over accuracy.

The criteria followed in this annotation guide aim to reduce the complexity for the machine of learning contextual nuances. It should be noted that no grammatical rules are involved in this functional class, so it is more evident that True Positives (TP) must be deduced from a broader context.

Besides, in some cases, we had many difficulties in agreeing to consider items as DM o non-DM because they did not appear in any DMs classification for Spanish nor English. Adverbs ending in *-mente* are one of those cases, as they should not be systematically considered DMs.

For instance, *especialmente* works as an adverb in some contexts (6):

(6) *En España destaca* ***especialmente*** *el negocio de Automóviles* (= ‘de manera especial’)

In other contexts, it has the function as a DM of highlighting (7), and it can be rephrased as a quantity adverb, standing out a member of discourse:

(7) *En nuestro caso la ejecución ha sido* ***especialmente*** *difícil* (= ‘muy’)

Another example of ambiguous DM is *con respecto al*, that functions as DM introducing a topic when it is at the beginning of a sentence (8), but not when it has a comparison function (9) or when it is in the middle of the sentence (10). We decided to annotate *con respect al* only when it appears at the beginning of a phrase or paragraph introducing a new idea:

(8) ***Con respecto a*** *la tecnología, estamos invirtiendo fuertemente para ser más eficientes y abaratar los procesos (DM)*

(9) *Ha mantenido el volumen de actividad* ***con respecto al*** *año anterior* (no DM)

(10*) En atención al compromiso adquirido hace un año* ***con respecto al*** *cumplimiento de todas las recomendaciones* (no DM).

The guidelines (available for consultation here[1]) are organised following three types of criteria: General criteria (general rules), inclusion criteria (positive rules) and exclusion criteria (negative rules). Furthermore, there is a section where all DMs annotated in the corpus are collected (288 in total).

The most relevant criteria are provided in the following paragraphs.

General criteria: As a general rule, we annotate discourse markers included in general classifications and others that do not appear in taxonomies. Still, we have added some DMs typical of the financial language (*adicionalmente*). We annotate all the words that are part of a discourse marker: those which include prepositions and articles, as *además*/*además* ***de/*** *además* ***del***; or those followed by a nexus: *de tal forma* ***que***. No punctuation marks are incorporated in the annotations. Discourse markers are collected without commas or dots. As an exception, a comma should be included following the marker in enumeration with ordinals to avoid the ambiguities caused by these elements functioning as determinative adjectives (*primero, segundo*).

Inclusion criteria: DMs included in the annotation guide follow the three classification criteria established by researchers: a) semantic criteria, since their inferences help us to group them into homogeneous types; b) syntactic criteria, because these characteristics let us limit the DMs when they are part of a larger constituent; and c) morphological criteria, related to their nature, form and grammaticalisation process. The inclusion criteria of a DM have also been decided according to the given contexts.

Exclusion criteria: The main principle in the negative rules is that we do not include DMs with a low degree of grammaticalisation[2]. Besides, we have not annotated those items whose form is identical to DMs, but which function as modifiers in other parts of the speech. As regards to specific negative rules, we are not including metatextual or anaphoric markers (*en este contexto, a partir de ahí, sobre esta base, hasta el punto de, dicho lo cual, centrándonos en*, etc.); only *todo ello,* considering its degree of gramaticalization. Due to their variability, some DMs, particularly those which are addressed to the audience, have multiple combinations: *como ven, como bien conoce, no cabe ninguna duda de que*, etc., or others alluding to personal opinions: *en nuestro caso, a mi juicio, a nuestro juicio, en mi opinión*, etc. In these cases, we do not include them either.

Another negative rule is not annotating discontinuous discourse markers: *no solo… sino también* or comparative structures: *tan… como, más…que*, etc. We also do not annotate markers that incorporate an element that modifies only part of the marker, not the whole marker. This means that the DM has a small degree of grammaticalisation or is not grammaticalised in that example, so such cases are not included: ***gracias****, en cierta medida,* ***a; con el objetivo*** *claro* ***de***. However, they are not exceptions to those particles that could have two parts (discontinuous DMs): *por un lado… por otro; por una parte… por otra*, since they can work independently from the other part (we can only have *por un lado*, or *por otra parte*, and they are doing a function by themselves).

3.2 Manual Annotation

This process was divided into three phases:

a. Training the two annotators with the guide and the tool (Doccano[3]). In this phase, both annotators could consult each other's annotations to reach a consensus and prove they had acquired the required skills. This process helped to modify some definitions in the annotation guide. In total, 100 LTS were annotated, one-quarter of the dataset.
b. Creation of the Gold Standard (GS). Each linguist annotated 40 LTS in an utterly blind way (i.e. without knowing the annotation of the other linguist and

[1]http://www.lllf.uam.es/ESP/Publicaciones/guia_anotacion.html

[2] The grammatizalisation process consists in the acquisition of a new grammatical value for these lexical units, which implies a shift from a more referential meaning to a less referential one. For instance, *es más* does not mean the beginning of a comparison structure, if not it appears alone in the speech guiding an inference reinforcing the following information. Compare: ***Es más****, en el siglo XXI en el que ya nos adentramos, el avance cada vez más rápido de la tecnología en combinación con la gestión más profesionalizada de la economía // nuestro mínimo regulatorio* ***es más*** *bajo porque nuestro modelo está menos interconectado y es más fácil de resolver*.

[3] https://doccano.herokuapp.com/

consulting only with the guide). This part is the one that has been used to calculate IAA (see 3.3.). The GS has been generated by joint approval of the two annotators after knowing the IAA results. It was not necessary for a judge to decide discrepancies. A first DM tagger has been created with the 100 + 40 LTS.

c. Manual revision of the automatic tagging generated by the initial DM tagger model. Each annotator has post-edited 130 LTS and corrected the assigned tags per tagger. Each annotator has acted as an expert judge in deciding the final version. There is no cross-checking between annotators. The result is a Silver Standard (SS) of 260 LTS.

The DM tagger has been trained on the first and third datasets, leaving the GS for evaluation (see 5).

3.3 Interannotator Agreement (IAA)

The inter-annotator agreement (IAA) measures how well different annotators can make the same annotation decision for a specific category. IAA also reveals how clear the annotation guidelines are and how reproducible the annotation task is. Cohen's kappa coefficient (κ) is a statistic to measure the reliability between annotators. It is more robust than the simple per cent of agreement (or accuracy) since κ considers the possibility of agreement by chance: $\kappa = (P_o - P_e) / (1 - P_e)$ where P_o is the relative observer agreement among annotators and P_e is the probability of agreement by chance.

Two annotators, we will refer to as A and B, worked with 40 documents, accounting for 52,890 tokens (words and punctuations) in 1,759 sentences. Annotators A and B recognised, respectively, 850 and 756 discourse markers agreeing in 732 cases (annotator A identified 118 cases not recognized by B, and B 33 cases not recognized by A). The IAA computed with κ was 0.897, which can be interpreted as a remarkably high degree of agreement.

Annotators went back to agree on their disagreements to build a reliable set to measure the classifier's performance. The number of discourse markers finally agreed was 856. The comparison between the original annotations and the new agreed set, calculated with κ, were 0.957 for annotator A and 0.916 for B. As human classifiers, the performance of the annotators is shown in Table 1, and it will be used as a reference when evaluating the performance of an automatic classifier. The formulas used to calculate this performance of a classifier are precision = TP / (TP + FP), recall = TP / (TP + FN), and F1-score = 2 * precision * recall / (precision + recall), where TP are the number of true positives, FP the number false positives, and FN the number of false negatives. We used seqeval (Nakayama, 2018) to calculate them.

Annotator	κ	Precision	Recall	F1
A	0.957	0.955	0.949	0.952
B	0.916	0.970	0.867	0.915

Table 1: Annotator performance on the test set after agreement.

4 Model Training and Selection

We used pre-trained transformer-based language models to approach the problem of discourse marker detection as a token classification task with a IOB (Input-Outside-Beginning) annotation scheme and one category (DM).

The annotated data was split into the training and validation sets. This data was annotated before the IAA experiment. The data finally used in the experiments are shown in Table 2.

Set	Sentences	Tokens	DMs
Training	3,735	118,406	1,880
Validation	938	30,524	440
Test	1,759	52,890	856

Table 2: Annotated sets.

We experimented with BSC-BNE[4], a Spanish Roberta model (Gutiérrez-Fandiño et al., 2021), mBERT[5] (Devlin et al., 2019) and BETO[6] (Cañete et al., 2020), and XLM-Roberta[7] (Conneau et al., 2020).

[4]https://huggingface.co/BSC-TeMU/roberta-base-bne

[5]https://huggingface.co/bert-base-multilingual-cased

[6]https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[7] https://huggingface.co/xlm-roberta-base

Model	STL	LR	Epochs	BS	WU	Avg. F1
BSC-BNE	all	5e-5	3	8	0.0	0.928
BETO	all	5e-5	4	8	0.0	0.927
mBERT	first	7e-5	4	16	0.1	0.927
BETO	first	6e-5	4	8	0.1	0.926
BETO	all	6e-5	3	8	0.1	0.926

Table 3: Models performance on the validation set, where STL is the sub-token labelling strategy (first or all), LR is the learning rate, BS is the batch size, and WU is the warmup ratio.

Following the recommendations in the Appendix A.3 for fine-tunning mBERT models (Devlin et al., 2019), we performed a grid search of hyperparameters for each language model with learning rates: 2e-5, 3e-5, and 5e-5; epochs: 2, 3, and 4; batch sizes: 8, 16, 32; warmup ratios: 0 and 0.1; and three different seeds (0, 3, and 5). We used the AdamW optimiser with no weight decay and a warmup of 0 and 10% steps. For each model, we have two versions: one that only labels the first sub-token delivered by the model's internal tokeniser and a second version where all the sub-tokens of a token are labelled. We averaged the F1-score of the three runs with different seeds to assign performance to a classifier. F1-scores were calculated with the SeqEval package[8] on a DM basis, i.e., a DM is correct if all the tokens in a DM have received the correct IOB tag.

The five best-performing systems, shown in Table 3, had a remarkably similar average F1-score, and the worst-performing system of the 2160 systems tested had a 0.882 F1-score[9].

5 Evaluation and error analysis

Finally, we trained a system with the model and the hyperparameters of the best performing system in Table 3 (with the seed set to 0), and it was evaluated with the test set. Results for this system were a precision of 0.941, a recall of 0.925, and an F1-score of 0.933, which is right in the middle of the range defined by the two human annotators (0.915–0.952).

Regarding false positives (FP) on the test set (38 cases in total), 40.67% of the cases were considered true false positives by the human annotators:

- 24.01% of the FP cases corresponded to ill-formed IOB sequences, mostly tokens labelled with inside tags (I-DM) without a beginning tag (B-DM) on the preceding token.
- 16.66% of the FP cases, despite being well-formed according to the OIB scheme, were considered valid false FP.

The rest of the FP cases (59.33%) were regarded as actual discourse markers, and they were distributed as follows:

- 42.59% of the cases were actual discourse markers that went unnoticed by the two annotators in the test dataset but were present in the training set and the annotation guideline.
- 3.7% were valid right-side extensions of other known discourse markers also present in the training set: *además de*, *en consecuencia de* or *de tal forma que*.
- 12.96% were accurate discourse markers, overlooked by the annotators without any occurrence in the training set. The tagger has been able to generalise that they can be DMs. These are the most interesting results, as they show that the model has been able to resolve doubts that arise for human annotators.

Some of these "new" discourse markers can be considered generalizations done by the model: *a continuación*, *al margen del*, *con ello*,*del mismo modo que*, *en total*, and *posiblemente*. Some of these particles were identified during the design of the guidelines. However, we did not annotate them in the GS as proper DMs because we considered them fuzzy. But the ML model has been able to learn in fuzziness.

On the other hand, False Negatives (FN) are DMs that were annotated by the linguists in the GS but were not detected by the ML model. In total, there were 52 FNs, of which:

- 75% (39 cases) were actual DMs. Hence model errors.

[8] https://github.com/chakki-works/seqeval

[9] Experiments with Bi-LSTM models hardly reach 70% F1-score on test set.

- 25% (13 cases) were genuine DMs, which the human annotators did not detect, and the model did.

In summary, the model has improved the performance of humans proportionally more on the FP side than on the FN side. Out of 90 cases, the ML model hits 22 (24.44%) versus human annotators.

6 Conclusions and future work

The proposed model shows an F1-score (0.933) in the range defined by two annotators (0.915–0.952), and error analysis of the false positives cases in the test set reveals that the model was able to recognise a significant number of discourse markers that went unnoticed by the annotators, some of them seen in the training set and a few discovered by the model.

In conclusion, we can say that DMs are units that mark the discourse and give it cohesion and coherence to facilitate the reader the comprehension and interpretation of the text. We also have concluded that they are an open or semi-open functional category. That is, they are not a grammatical category, although most of the DMs are in a grammaticalisation process. We know, for sure, that they are units of the speech that mark the discourse. So, which words can mark the discourse and perform the functions of a DM? The complexity of this DM-tagger and everything related to DMs is, precisely, that we are dealing with a difficult task for linguists to define.

We assume this work is challenging, and it wasn't easy to define the criteria for considering an element DM or no-DM. There were and still are some doubts about the definition and the limits of these discourse units. Overall, we had difficulties with DMs coming from adverbial phrases because their context tends to be ambiguous. The guideline of this study, especially its negative criteria, must be revised. Nevertheless, our model can discover or annotate new DMs that were not initially annotated by humans, which means that NLP can somehow develop the capacity of detection DMs functionality beyond their form.

Further work will be, indeed, a Discourse Marker Tagger that classifies DMs into their types and subtypes (following the work begun by Hernán and Nazar (2018) and Nazar (2021), section 1.3) because this would provide us more information about the financial text and its factual inferences. We will look to study derived from usage data, with less reliance on language knowledge, using the methodology proposed by these authors. These steps may make us closer to defining Discourse Markers better than we used to do through human introspection.

The DMs tagger (under development) will be used in the annotation of argumentative structures. In particular, we are mainly interested in CAUSE-EFFECT (This has happened. *Consequently*, this other thing has happened) and counter-argumentative structures (This has happened. *However*, this other thing has also happened).

A DM-tagger and its automatic annotation may be an objective measuring instrument to help resolve theoretical discussions.

Acknowledgements

The research has been carried out within the CLARA-FINT project (PID2020-116001RB-C31), funded by the Spanish Ministry of Science and Innovation.

References

Alonso, L., Castellón, I., Gibert, K. and Padró, L. 2002. Lexicón computacional de marcadores del discurso. Procesamiento del lenguaje natural, 29:239–246.

Bel, N., Bracons, G., and Anderberg, S. (2021). Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis. *Information*, *12*(8), 307. http://dx.doi.org/10.3390/info12080307.

Briz, A., Pons, S. and Portolés, J. (coords.) 2008. *Diccionario de partículas discursivas del español*. Retrieved from www.dpde.es.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In PML4DC at ICRL-2020.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, Veselin. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the ACL*.

Devlin, J., Chanh, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language

understanding. In *Proceedings of the NAACL*.

El-Haj, M., Rayson. P., Walker, M., Young, S., and Simaki, V. (2019). In search of meaning: lessons, resources, and next steps for computational analysis of financial discourse. *Journal of Business, Finance and Accounting*, *46*(3-4), 265-306. https://doi.org/10.1111/jbfa.12378.

Fuentes Rodríguez, C. 2009. *Diccionario de conectores y operadores del español*. Madrid: Arco Libros.

Gisbert, A. 2021. Financial narratives. In Moreno-Sandoval, (coord.), *Financial narrative processing in Spanish*. Valencia: Tirant lo Blanc, 15-50.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Pio-Carrino, C., Gonzalez-Aguirre, A., Armentano-Oller, C., Rodriguez-Penagos, C. and Villegas, M. 2021. Spanish Language Models. https://arxiv.org/abs/2107.07253.

Landone, E. 2012. La clasificación de los marcadores del discurso y su valor operativo. In *Cassol, A., XXIV Congresso AISPI*, 431-444. Roma: AISPI Edizioni.

Llamas Saíz, C., Martínez Pasamar, C., and Tabernero, Sala, C. 2012. La comunicación académica y profesional. Usos, técnicas y estilo. Pamplona: Thomson Reuters / Aranzadi (pp. 140-141).

Loureda, Ó. and Acín, E. (coords.) 2010. *Los estudios sobre marcadores del discurso en español hoy*. Madrid: Arco/Libros.

Martín Zorraquino, M.A. and Portolés, J. 1999. Los marcadores del discurso. In *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, 4051-4214.

Mateo Martínez, J. 2007. El lenguaje de las ciencias económicas. In E. Alcaraz, J. Mateo and F. Yus (Eds.), *Las lenguas profesionales y académicas* (pp. 191-203). Barcelona: Ariel.

Montolío, E. 2001. *Conectores de la lengua escrita. Contraargumentativos, consecutivos, aditivos y organizadores de la información*. Barcelona: Ariel.

Moreno-Sandoval, A., Gisbert, A., Haya, P.A., Guerrero, M. and Montoro, H. 2019. Tone analysis in Spanish financial reporting narratives. In M. El-Haj, P. Rayson, S. Young, H. Bouamor and S. Ferradans (Eds.), *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019*) (pp. 42-50). Turku: Linköping University Electronic Press.

Moreno-Sandoval, A. Gisbert, A. and Montoro, H. 2020. FinT-esp: a corpus of financial reports in Spanish. In Fuster-Márquez, Gregori-Signes and Santaemilia-Ruiz (eds.) *Muliperspectives in analysis and corpus design*. Granada: Comares, 89-102.

Muller, P., Conrath, J., Afantenos, S. and Asher, N. 2016. Data-driven discourse markers representation and classification. In *TextLink– Structuring Discourse in Multilingual Europe Károli Gáspár University of the Reformed Church*. Hungary, Budapest.

Nakayama, H. 2018. seqeval: A Python framework for sequence labeling evaluation. https://github.com/chakki-works/seqeval.

Nazar, R. (2021). Inducción automática de una taxonomía multilingüe de marcadores discursivos: primeros resultados en castellano, inglés, francés, alemán y catalán. In *Procesamiento del Lenguaje Natural*, (67), 127-138.

Pons, S. 2000. Los conectores. En A. Briz and Val.Es.Co, (eds.), *¿Cómo se comenta un texto coloquial?* Barcelona: Ariel, 193–220

Robledo, H., Nazar, R and Renau, I. 2017. Un enfoque inductivo y de corpus para la categorización de los marcadores del discurso en español. In *Proceedings of the 5th International Conference "Discourse Markers in Romance Languages: Boundaries and Interfaces"*, 91–93. Université Catholique de Louvain, Belgium.

Robledo, H. and Nazar, R. 2018. Clasificación automatizada de marcadores discursivos. In *Procesamiento del Lenguaje Natural*, (61), 109–116.

Vargas-Sierra, C. and Carbajo-Coronado, B. 2021. Anglicisms in Financial Narrative: The Case of the Annual Reports and Letters to Shareholders. In Moreno-Sandoval, (coord.) *Financial narrative processing in Spanish*. Valencia: Tirant lo Blanc, 99-134.